

# Modeling and Analysis of Cloud Signaling Services

Eduardo M. Hargreaves<sup>1</sup>, Paulo H. de Aguiar Rodrigues<sup>1</sup>, Daniel S. Menasché<sup>1</sup>

<sup>1</sup>Departamento de Ciência da Computação  
Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, Brasil

eduardo.hargreaves@ppgi.ufrj.br, aguiar@ufrj.br, sadoc@dcc.ufrj.br

**Abstract.** *Networks connecting distributed cloud services through multiple data centers are called cloud networks. These types of networks play a crucial role in cloud computing and a holistic performance evaluation is essential before planning a converged network-cloud environment. We analyze a specific case where some resources can be centralized in one datacenter or distributed among multiple data centers. The economy of scale in centralizing resources in a single pool of resources can be overcome by an increase in communication costs. We propose an analytical model to evaluate tradeoffs in terms of application requirements, usage patterns, number of resources and communication costs. We numerically evaluate the proposed model in a case study inspired by the oil and gas industry, indicating how to cope with the tradeoff between statistical multiplexing advantages of centralization and the corresponding increase in communication infrastructure costs.*

## 1. Introduction

Cloud computing is a large-scale distributed computing paradigm that is driven by economies of scale, in which a pool of computing resources (e.g., networks, servers, storage, applications, and services) are provisioned, delivered and released on demand to users over a network [1]. To the user, the available capabilities often appear to be unlimited and can be elastically provisioned in any quantity at any time [2].

To improve cloud performance and resiliency, the current trend is to deploy the services across multiple and geographically distant sites [3]. In this context, cloud networks (networks connecting cloud services hosted in multiple data centers) play a crucial role in cloud computing and are an indispensable ingredient for high performance cloud computing [4].

This work is motivated by a real-world project of deployment of an Oil & Gas (O&G) application in multiple sites. Because O&G possesses a high cost per license, buy one license per user can be very expensive. Because a user can run multiple applications, like e-mails clients or spreadsheets softwares, the single user licensing approach results in a waste of resources due to license underutilization. One solution often used in corporate environments for cost minimization is share a license between multiple users with the adoption of the floating licensing approach [5]. In this approach, a limited number of licenses is stored in a pool shared among a large number of users. When a user wishes to run the application, a license is requested to a license server. The task of a software license server is to determine and control the number of active replicas of the software based on the license entitlements that an organization owns or the system capabilities. If a license is available, the license server allows the application to run. When the user

finishes the application, the license is reclaimed by the license server and made available to other users. To avoid user's misbehavior, the application periodically sends license renewal requests to the license server. If the client machine does not receive an answer, a timeout is detected and the application is closed.

The sharing of scarce resources (licenses) among users taking advantage of the probabilistic pattern of user's access produces *statistical multiplexing gains*. The sharing increases the utilization of the licences and reduces the number needed to satisfy users requirements. This gain can be viewed as a consequence of the law of large numbers and increases as the number of users sharing a resource increases. It is important to notice that there are more users than licenses, therefore it is possible that a user cannot get an available license. In this case he waits or he is blocked. Systems of the first type are called *waiting systems* while the second type systems are called *loss-systems* [6]. We modelled the licensing system as a *loss-system*. The probability of not get a license is the *blocking probability*. Because blocking is inconvenient to users, the blocking probability is a Service Level Agreement (SLA) parameter.

A cloud is as a set of one or more *resource units (RUs)* running services that are requested by users. Application server and licence server are examples of servers running in RUs. The set of *O&G* applications is so image intensive and demands so much bandwidth that users and RUs running the application servers have to be in the same local area network due to performance constraints. The cloud architecture can be a federated network of public or private clouds, hosting the software floating license cloud service, the *O&G* application itself and other corporate applications. Licenses are consolidated remote to users in a single pool (Figure 1) or near to users spread over multiple pools in multiple sites (Figure 2).

The single pool and multiple pool architectures considered in this paper make use of soft-state signaling solutions [7]. In soft-state solutions, installed state at a remote site needs to be frequently refreshed by clients running the application. Otherwise, the state times out (and is removed). Soft-state solutions are quite common, being used in a variety of protocols as RSVP, SRM, SIP and IGMP, to name a few. In hard-state solutions, in contrast, state remains installed until explicitly removed. Soft-state requires well-provisioned links for frequently updating states at remote sites through a WAN.

In the single pool case, the federated clouds are linked by the network cloud in such a way that background traffic competes with the request/response signaling traffic. Due to congestion generated by the background traffic, timeouts can occur and the application might be prematurely closed. In this situation, the cloud network capacity needs to be increased to reduce the premature timeout probability. In the multi-pool case, users access local (but typically more scarce) resources at distributed pools, the connection is typically over over-provisioned local area networks (LANs) and timeouts due to congestion rarely occur. While the single pool explores the statistical multiplexing gains due to resource sharing, such economy of scale gains might be precluded by an increase in communication costs.

1. How to quantify advantages and disadvantages of consolidating resources in a single pool?
2. What is required (in terms of resources and infrastructure upgrades) to satisfy the service level agreements?

While partially answering the questions above, our key contributions are the following:

1. **holistic analysis of license serving:** we propose an integrated framework for the assessment of benefits and costs of service infrastructures accounting for software and network aspects. We apply our framework to the analysis of how to distribute resources in a cloud environment;
2. **analytical model:** we specialize the proposed framework to analyze the tradeoffs involved in resource distribution accounting for gains due to statistical multiplexing and costs due to congestion;
3. **case study:** we numerically investigate the proposed model using a case inspired by a real-world Oil & Gas setup.

The remainder of this paper is organized as follows. Related work is discussed in Section 2. In Section 3 we present the proposed analytical model. Section 4 contains the numerical examples, Section 5 presents further discussion about the model applicability and Section 6 concludes the paper.

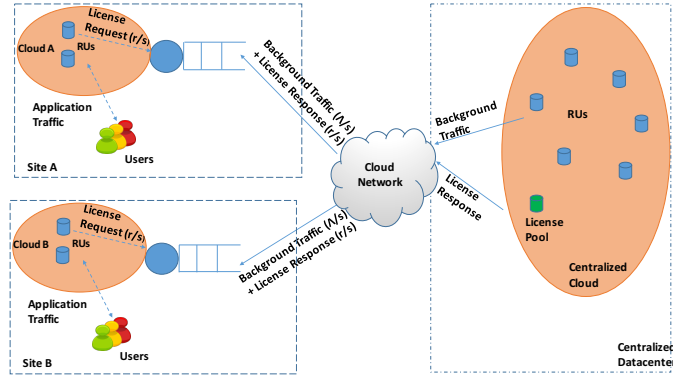
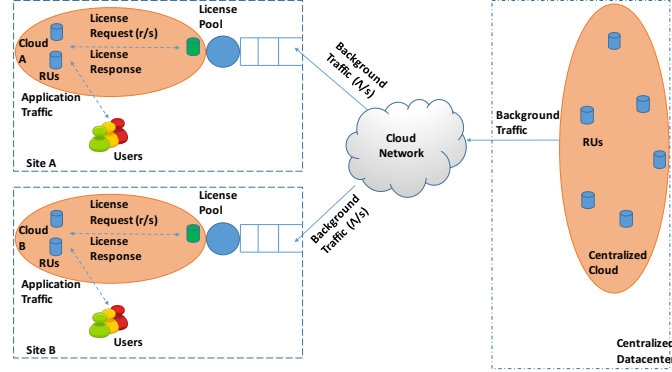


Figure 1. Centralized servers (single pool)

## 2. Related Work

Since networking has a strong impact on end-to-end cloud service, a holistic performance evaluation is essential before planning a converged network-cloud environment. Performance evaluation on the sufficient number of computational resources necessary to meet a desired Service Level Agreement (SLA) in a cost-effective way has attracted extensive research interest [8, 9]. [10] has proposed a modeling and analysis approach by exploiting network calculus theory to define a general profile that can characterize service capability of either a network or Cloud service.

Our work is closely related to the facility location problem in operations research [11]. In facility location problems, the aim is optimally place facilities so as to minimize transportation costs. While we also target minimization of service costs, we



**Figure 2. Distributed servers (multiple pools)**

consider the interplay between those and statistical gains due to multiplexing, timeout probabilities due to congestion and blocking probability due to resource exhaustion.

[9] modeled the blocking probability of a cloud with a large number of Resource Units (RUs) and general service times given by a  $M/G/m/m + r$  queuing system with single task arrivals and a task buffer of finite capacity. [8] evaluated finite population and heterogeneous resource requests using the blocking probability as an SLA performance measure to dimension clouds. They showed that infinite source models may lead to an overestimation of the number of RUs.

Although the papers above are related to ours, none of them considered the trade-off between gains due to statistical multiplexing and costs to support remote traffic. While [9] and [8] considered blocking of RUs, [10] considered network-cloud service capabilities. To the best of our knowledge, this paper is the first to bridge the two aspects in an integrated manner.

### 3. Model and Notation

We begin by describing the system of interest in this paper. A cloud is a pool of one or more RUs that can be requested by users. These units can be virtual machines (VM) or CPUs. We assume that there are RUs acting as signaling (control) servers and application servers. Each signaling server controls one pool of application states. To simplify presentation, we assume that each running instance of the application is associated to an application state.

We consider a population of  $S$  users divided into  $N$  sites (or pools in the distributed case). Each site  $i$  comprises  $S_i$  users. Each user initiates new application instances at rate  $\lambda$ . If the signaling server is able to allocate space to store state information associated to the new instance, the requester accesses the VM until the service is completed. When the user finishes its session, the space used to store state information is reclaimed by the license server and made available to other users. Each session duration is exponentially distributed with mean  $1/\mu$  hours. Let  $\rho = \lambda/\mu$ . Table 1 summarizes the

variable	description
$N$	number of sites
$S$	population size
$S_i$	population at site $i$
$L$	number of supported application states (resources)
$L_i$	number of supported application states (resources) at pool $i$
$\lambda$	request arrival rate per user (requests/hour)
$1/\mu$	average duration of a session (hours)
$\rho$	$\lambda/\mu$
$C_0$	initial circuit capacity (Mbps)
$C'$	additional circuit capacity (Mbps)
$C$	circuit capacity (Mbps) ( $C = C' + C_0$ )
$1/M$	packet mean size (bits <sup>-1</sup> )
$\Lambda$	background packet arrival rate (pkts/s)
$r$	rate at which application states are checked (s <sup>-1</sup> )
$\tau$	timeout detection threshold (s)
metric	description
$b_c$	blocking probability in the centralized setup
$b_d$	blocking probability in the distributed setup
$p$	timeout probability in the distributed setup
$s_c$	success probability in the centralized setup
$s_d$	success probability in the distributed setup
$s$	required success probability defined in SLA

**Table 1. Table of notation**

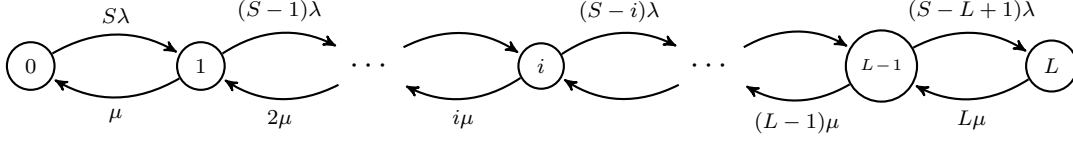
notation used throughout the rest of the paper.

Based on observed real-world signaling packets, application periodically send state renewal requests to the signaling servers at fixed rate of  $r$  attempts/s and detects a timeout if it does not receive an answer after  $\tau$  seconds. In the single pool architecture requests traverse a circuit with capacity  $C$  competing with background traffic. Background packets arrive at rate  $\Lambda$  pkts/s and each packet requires exponentially distributed service with mean  $1/M$ .

### 3.1. Blocking probability

To characterize the blocking probability, we use a finite source queueing model with homogeneous service distribution and no waiting time, as given by the well-known *Engset formula* [6]. The aggregated arrival rate of renewal requests is proportional to the number of idle users and the aggregate service rate is proportional to active users to whom access was granted. The state transition diagram of the finite source model is illustrated in Figure 3.

The states of the application replicas can be consolidated in a single pool or distributed across multiple pools. Let  $b(L, S)$  be the blocking probability for a population of



**Figure 3. State transition diagram for the finite source model**

size  $S$  associated to a signaling server that has capacity to handle  $L$  application states.

$$b(L, S) = \frac{\binom{S-1}{L} \left(\frac{\lambda}{\mu}\right)^L}{\sum_{i=0}^L \binom{S-1}{i} \left(\frac{\lambda}{\mu}\right)^i} \quad (1)$$

The numerical calculation of (1) leads to numerical problems for large values of  $S$  and  $L$ . So, to compute  $b$ , we used the following numerically stable recursive formula [6]:

$$b(j, S) = \frac{\rho(S-j+1)b(j-1, S)}{j + \rho(S-j+1)b(j-1, S)}, \text{ where } b(0, S) = 1 \text{ and } 1 \leq j \leq L \quad (2)$$

When referring to the centralized setup, we might denote  $b(L, S)$  simply by  $b_c$ . In the centralized case,  $b_c$  is given by:

$$b_c = b(L, S) \quad (3)$$

In the distributed scenario, there is one signaling server at each pool. Users compete for access to the signaling server of their corresponding pools and each pool  $i$  has its associated blocking probability  $b(L_i, S_i)$ . Let  $b_d$  be the blocking probability in the distributed case, given as the weighted sum of the blocking probabilities at each pool. So, the blocking probability in the distributed case is given by:

$$b_d = \sum_{i=1}^N \frac{S_i}{S} b(L_i, S_i) \quad (4)$$

where

$$S = \sum_{i=1}^N S_i, \quad L = \sum_{i=1}^N L_i \quad (5)$$

### 3.2. Timeout Due To Congestion

In this section we assume that each client periodically sends state renewal requests to the signaling server, at a constant rate of  $r$  requests per second. Each renewal request is also referred to as a *probe*. If the client does not receive an answer in  $\tau$  seconds ( $\tau < 1/r$ ), a timeout is detected and the application is closed. Note that due to network congestion, the chances of occurring a timeout if the signaling server is centralized is higher than if the signaling server is closer to end-users.

Let  $D_i$  be the duration of the  $i$ -th subinterval in which the state is checked. Let  $\tilde{N}$  be the number of times that probes are sent to the server until a timeout is generated,

without counting the last probe ( $\tilde{N} \geq 0$ ). Let  $q$  be the probability that a probe successfully yields a state renewal. Then,  $\tilde{N} \sim \text{Geometric}(1 - q)$ ,

$$P(\tilde{N} = n) = q^n(1 - q), \quad n \geq 0 \quad (6)$$

Let  $D$  be the time until a timeout, where  $D$  is given as a function of  $D_i$  and  $N$  as follows,

$$D = D_0 + D_1 + D_2 + \dots + D_{\tilde{N}} + D_{\tilde{N}+1} \quad (7)$$

where  $D_0 = 0$ . In the remainder of this section we assume that  $D_i \sim \text{Constant}(\frac{1}{r})$  for  $i = 1, 2, 3, \dots$  but the argument presented below easily generalizes for the case where  $D_i$  has general distribution, and  $D_1, D_2, \dots$  are independent and identically distributed with Laplace transform given by  $e^{(-s/r)}$ .

Conditioning on  $\tilde{N}$ , the Laplace transform of  $D$  can be derived as:

$$D^*(s) = \tilde{N}(z) \Big|_{z=e^{-(\frac{s}{r})}} D_{\tilde{N}+1}^*(s) \quad (8)$$

where  $\tilde{N}(z)$  is the z transform of  $\tilde{N}$ ,

$$\tilde{N}(z) = \frac{1 - q}{1 - zq} \quad (9)$$

and

$$D_{\tilde{N}+1}^*(s) = D_i^*(s) = e^{-s/r}, \quad i = 1, 2, 3, \dots \quad (10)$$

We substitute (9) and (10) into (8) to obtain

$$D^*(s) = e^{-s/r} \left( \frac{1 - q}{1 - e^{-s/r}q} \right) \quad (11)$$

Next, we characterize the network congestion which will cause the timeouts. This characterization is the first step in deriving a formula for the timeout in terms of  $C$  in such a way that tradeoffs can be evaluated in terms of  $L$  and  $C$ .

Let  $Y$  be the delay experienced by a probing packet that traverses a link with capacity  $C$  subject to exogenous (background) traffic with arrival rate  $\Lambda$  and size exponentially distributed with mean  $1/M$ . The CDF of  $Y$  is given by  $F_Y(y) = P(Y < y)$ . Then, the probability that a state renewal request succeeds is given by  $q = P(Y < \tau)$ . In the remainder of this paper, we assume that the delay is characterized by an M/M/1 queue, and that the overhead caused by state renewal requests into the link is negligible compared to the exogenous traffic. Then, it follows from [12, equation (5.119)] that

$$q = 1 - e^{-(MC - \Lambda)\tau} \quad (12)$$

The *target session duration* is the desired duration of the session as determined by the user requirements. Let the target session duration time be exponentially distributed with mean  $1/\mu$ , and let  $L$  be the random variable that characterizes the target session duration,  $L \sim \text{Exponential}(\mu)$ . The actual session duration might be smaller than the target session duration if a timeout occurs.

Let  $p$  be the probability of timeout. A timeout occurs if the target session duration is greater than the time until a renewal request fails. Therefore,  $p$  is given by:

$$p = P(L > D) = D^*(s) \Big|_{s=\mu} = D^*(\mu) \quad (13)$$

where  $D^*(\mu)$  is the Laplace transform of  $D$  given by (8), evaluated at the point  $s = \mu$ .

Our goal is to obtain a simple expression for  $p$  as a function of  $C$ . To this aim, we substitute (12) into (11) and make use of (13) to obtain  $p$ ,

$$p = e^{-\mu/r} \left( \frac{e^{-(MC-\Lambda)\tau}}{1 - e^{-\mu/r} (1 - e^{-(MC-\Lambda)\tau})} \right) \quad (14)$$

Note that if  $q = 0$  then  $p = e^{-\mu/r}$ . This is an expected result, as  $q = 0$  means that the state renewal will fail and timeout will only occur if the target duration is greater than  $1/r$ , the interval between renewal requests. The probability that the application cannot be started at first place is captured through  $b_c$ .

After some algebraic manipulation, it is possible to derive an explicit formula for the required link capacity  $C$  as a function of the targeted timeout probability  $p$ , the application characteristics ( $\tau$  and  $1/r$ ), usage patterns ( $\mu$ ) and background traffic characteristics ( $\Lambda$  and  $M$ ):

$$C = \frac{1}{M} \left( \Lambda + \frac{1}{\tau} \ln \left( \frac{p(e^{\frac{\mu}{\tau}} - 1)}{p - 1} \right) \right) \quad (15)$$

### 3.3. Success Probability

We consider the probability of success  $s$  the main SLA parameter. The success is defined by being granted access to the application and being successful in all attempts of state renewal.

Let  $s_c$  be the probability of success in the centralized setup. This probability is the product of the probability of not being blocked ( $1-b_c$ ) times the probability that all attempts succeeded ( $1-p$ ), with  $b_c$  and  $p$  given by (1) and (14).

$$s_c = (1 - b_c)(1 - p) \quad (16)$$

Let  $s_d$  be the probability of success in the distributed scenario. As the application and the signaling server are in the same cloud  $p$  is zero. On the other hand, the signaling servers at each pool have less resources than the central server, which might increase the probability of blocking. Then,  $s_d$  is given by:

$$s_d = 1 - b_d \quad (17)$$

In light of equations (16) and (17), we are ready to quantify the tradeoff mentioned in the beginning of this paper. The centralized setup is associated to a smaller blocking probability as resources are multiplexed in a single pool. Nonetheless, users incur a timeout probability due to network congestion. This tradeoff motivates an optimization problem, where the network designer is faced with a decision between centralizing the pool of resources or distributing resources across multiple pools.



### 3.4. Optimization Problem

The optimization problem consists of minimizing costs for a given success probability defined in an SLA. Formulas (1), (4) and (14) are consolidated in equations (16) and (17). These equations give the success probability in terms of application usage patterns, number of users, licenses and pools, renewal attempts characteristic, capacity and traffic of the network cloud. The communication and resource costs determine the merits of the centralized and distributed scenarios.

Next, we present the resource allocation problem in the centralized scenario:

$$\text{minimize : } c = \alpha L + \beta n C' \quad (18)$$

$$\text{subject to : } s_c \geq s \quad (19)$$

$$\text{constraint on variables : } L \geq 0, C' \geq 0 \quad (20)$$

The corresponding distributed resource allocation problem is:

$$\text{minimize : } c = \alpha L + \beta n C' \quad (21)$$

$$\text{subject to : } s_d \geq s \quad (22)$$

$$\text{constraint on variables : } L \geq 0, C' = 0 \quad (23)$$

where  $c$  is the cost in the centralized or distributed scenario,  $\alpha$  is the cost per maintained state (resource),  $\beta$  is the cost per Mbps and  $n$  is the number of sites that need to be upgraded in terms of capacity. We consider a link with initial capacity  $C_0$ . Let  $C'$  be the marginal capacity added to the link. Let  $C = C_0 + C'$  be the total link capacity (also referred to simply as *link capacity*). Note that the timeout probability  $p$  is a function of the total capacity  $C$ . The probability of success is given by (16) and (17) in the centralized and distributed scenarios, respectively. The scenario with the minimum cost between the centralized resource allocation problem and distributed resource allocation is the scenario with the total minimum cost.

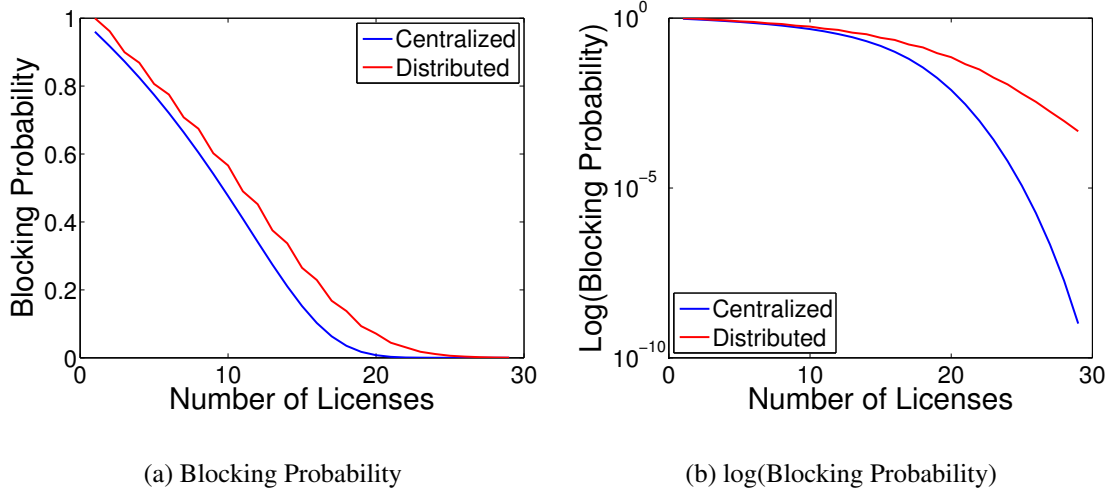
Let  $c_c^*$  and  $c_d^*$  be the minimum costs achieved through the centralized and distributed setups, respectively. Then, the network designer chooses the centralized or the distributed setups so as to minimize the minimum feasible cost  $c^*$ :

$$c^* = \min(c_c^*, c_d^*) \quad (24)$$

In order to find the optimal solution, we used an *interior point method* in which the derivatives are approximated by a solver as implemented in Matlab<sup>®</sup> convex optimization toolbox.

## 4. Numerical Examples

In this section we numerically investigate the proposed model. Our goals are to a) numerically illustrate the tradeoffs between blocking probability and timeout probability and b) indicate the applicability of the optimization problem proposed, quantifying the advantages and disadvantages of central and distributed pools of resources. Our examples in this section are motivated by the previously mentioned Oil & Gas application. In this



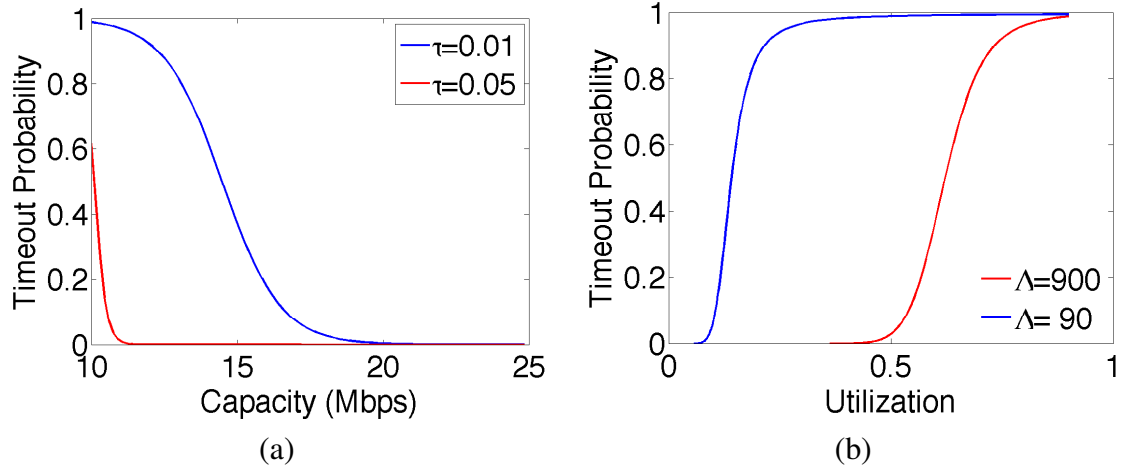
**Figure 4. Blocking probability,  $\frac{1}{\mu} = 8$  hours,  $S = 30$  users,  $\rho = 0.8$ , blue=centralized, red=distributed versus number of licenses**

application, signaling servers associate one license to each running instance of an application. Therefore, in what follows we refer to *licenses* and *application state resources* interchangeably.

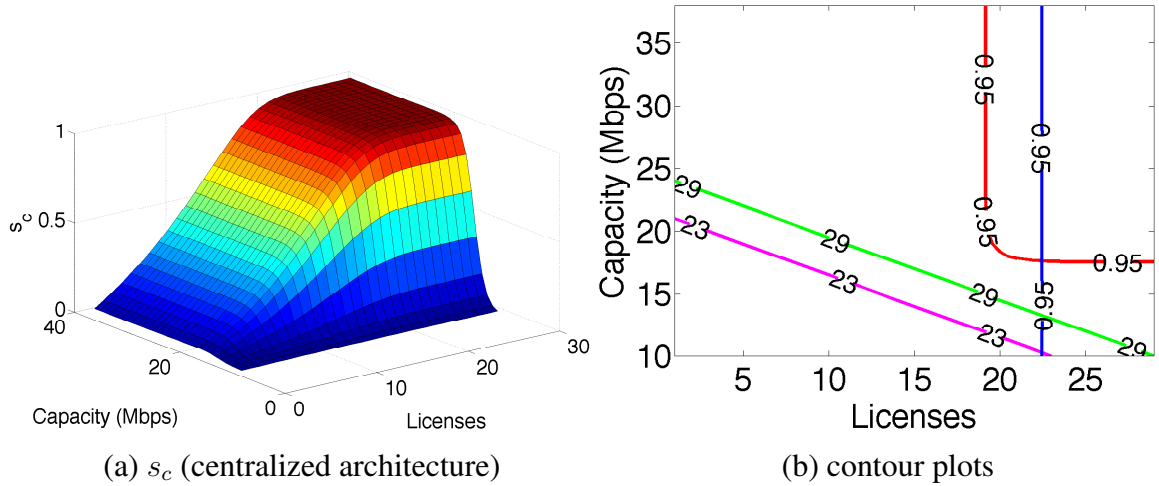
The first analysis is a comparison between the blocking probability in the centralized and in the distributed case as shown in Figure 4(a). In terms of blocking probability, the centralized architecture is always better. The logarithm of the blocking probability, plotted in Figure 4(b), shows that the advantage of the centralized architecture increases as the number of available licenses increases. We refer to the reduction in blocking probability due to centralization as *licensing statistical multiplexing gains*.

Figure 5(a) shows the timeout probability due to congestion  $p$  as a function of the link capacity  $C$  for  $\tau = 0.01$  and  $\tau = 0.05$ . As the capacity  $C$  increases,  $p$  decreases. When  $C = \infty$  (over-provisioning) we have  $p = 0$ . Figure 5(a) also shows the significant impact of the timeout detection threshold  $\tau$ . When  $\tau = 0.05$ , a small increase in capacity can reduce  $p$  to zero. When  $\tau = 0.01$  it is necessary to double the capacity (from 10 Mbps to 20 Mbps) to achieve the same result. In as  $\tau$  increases,  $p$  sharply decreases to zero for small values of  $C$ . On the other hand, when  $\tau \approx 0$  the centralized architecture is infeasible. Roughly speaking, if an application developer wishes to allow the application to be used in different sites with a centralized floating licensing approach and the network capacity is small, the parameter  $\tau$  must be relaxed.

Figure 5(b) shows how the timeout probability due to congestion ( $p$ ) varies as a function of the utilization factor for different workloads and link capacity ranges. We vary the link capacity in the range of [10,25] (red line) and [1,15] (blue line) adjusting the offered workload accordingly. Given a target value for  $p$ , the higher capacity network (red line) can operate at a higher utilization level than the lower capacity network (blue line). We refer to this increase in supported utilization due to increased capacity as *networking statistical multiplexing gain*. One of its consequence is that it is better to have one higher-speed link instead of having  $n$ -parallel lower-speed links to carry the same amount of traffic [13].



**Figure 5. Timeout probability due to congestion ( $p$ ) for different values of (a)  $\tau$  (for fixed  $\Lambda = 900$  pkts/s) and (b)  $\Lambda$  (for fixed  $\tau = 0.01$ s).  $\frac{1}{\mu} = 8$  hours,  $\frac{1}{r} = 120$  s,  $\frac{1}{M} = 1250$  Bytes. In figure (b): red line,  $\Lambda = 900$  pkt/s and  $C \in [10, 25]$  Mbps; blue line,  $\Lambda = 90$  pkt/s and  $C \in [1, 15]$  Mbps.**



**Figure 6. Graphical solution of optimization problem: (a)  $s_c$  as a function of capacity and number of licenses, (b) contour plots of  $c_c = L + 2C'$  (magenta and green lines),  $s_c = 0.95$  (red curve) and  $s_d = 0.95$  (blue curve).  $\frac{1}{\mu} = 1/\lambda = 8$  hours,  $S = 30$  users,  $\frac{1}{\mu} = 8$  hours,  $1/r = 120$  s,  $\tau = 0.01$  s,  $1/M = 1250$  Bytes,  $\Lambda = 900$ ,  $C_0 = 10$  Mbps. “Capacity” refers to  $C = C_0 + C'$ .**

Figure 6(a) shows the centralized probability of success  $s_c$  as a function of the number of available licenses and link capacity. The success probability approaches one when both the link capacity and the number of licenses are increased. Note that unilaterally over-provisioning the link capacity or the number of licenses is not sufficient in order to achieve high values of  $s_c$ .

The optimization problem described in Section 3.4 admits a graphical solution. We start considering the centralized architecture. Recall that  $C'$  is the marginal capacity added to a link of initial capacity  $C_0$  resulting in a link with capacity  $C = C' + C_0$ . Varying the values of  $C'$  and  $L$ , we affect the cost given by (18). For each value of the cost variable,

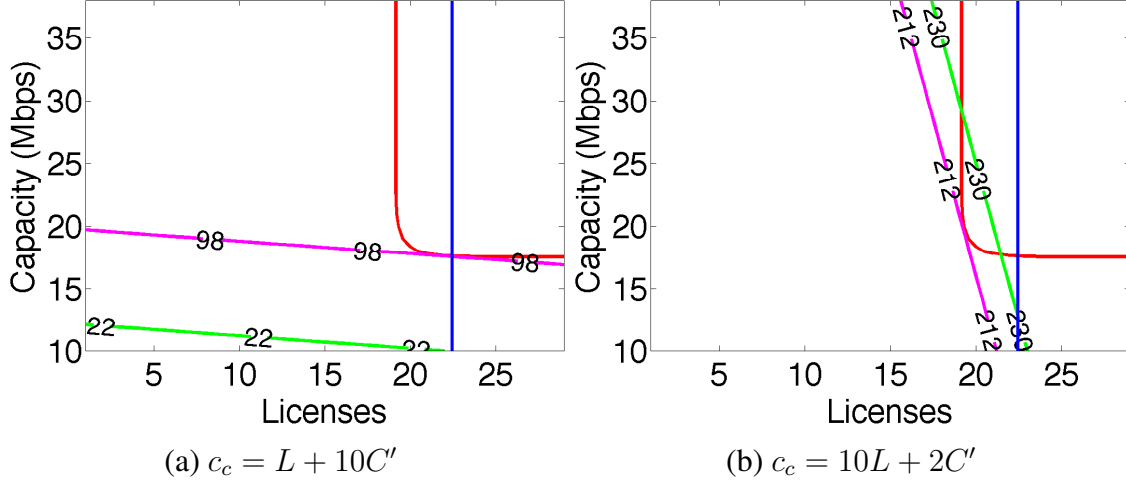
$c$ , we have a corresponding line in the  $(L, C)$  plane characterized by  $c = \alpha L + \beta n(C - C_0)$ . Given a set of cost values, we characterize a family of parallel lines in the  $(L, C)$  plane, where  $L = 1, 2, \dots$  and  $C \geq C_0$ . The minimum value of  $c$  for which the corresponding line intersects the curve associated to the constraint  $s_c \geq s$  corresponds to the optimal centralized solution  $c_c^*$ . In the distributed architecture, let  $L^*$  be the minimum value of  $L$  for which the constraint  $s_d \geq s$  is satisfied. As in the distributed architecture we assume that licenses and additional state resources are accessed locally, we have  $C' = 0$ . Therefore, the optimal distributed solution  $c_d^*$  occurs at  $(L, C) = (L^*, C_0)$ . We compare the centralized and distributed solutions and select the one with minimum cost.

Figures 6(b), 7(a) and 7(b) illustrate the graphical solution. In all cases the population consists of 30 users,  $C_0 = 10$  Mbps and  $s = 0.95$ . When considering distributed solutions, the population is split equally between two distinct sites. In Figure 6(b) the cost per license is twice the cost per Mbps, meaning that  $L + 2C' = c$ . The cost curve marked with “23” (magenta line) represents a scenario in which  $c = 23$ , i.e.,  $L + 2C' = 23$ . The cost curve marked with “29” (green line) represents a scenario in which  $c = 29$ , i.e.,  $L + 2C' = 29$ . The red curve is a contour plot of the centralized architecture constraint wherein  $s_c = 0.95$ , and the blue curve is a contour plot of the distributed architecture constraint wherein  $s_d = 0.95$ . The intersection between a cost line and a constraint curve corresponding to the smallest feasible cost occurs at the bottom of Figure 6(b). This means that the distributed architecture is the best choice,  $c^* = c_d^* = 23$ , and  $L = 23$  is the minimum number of licenses that satisfies the SLA requirement. To achieve the same SLA, the cost of the centralized architecture would be 29.

Figure 7(a) shows the graphical solution of the optimization problem when the cost per Megabit/s is five times the cost per license. In this case the distributed architecture is again the best solution. The intersection between a cost line and a constraint curve corresponding to the smallest feasible cost occurs at the bottom of Figure 7(a). At point  $(L, C) = (22, 10)$  (or, equivalently,  $(L, C') = (22, 0)$ ), the cost line  $L + 10C' = c_d^*$  (magenta line), where  $c^* = c_d^* = 22$ , intersects the distributed architecture constraint given by  $s_d = 0.95$  (blue line). In this case, to satisfy the service level agreement the cost of the centralized architecture would be  $c_c^* = 98$ . The communication cost makes the centralized architecture not viable.

Figure 7(b) shows an example where the centralized architecture outperforms the distributed one. The cost per license is five times the cost per Megabit/s. In this case,  $10L + 2C' = c_d^*$  (magenta line), where  $c^* = c_d^* = 212$ , intersects the red curve associated to the centralized architecture constraint,  $s_c = 0.95$ . The blue curve corresponding to the distributed architecture constraint does not intersect lines associated to costs smaller than or equal to 230.

As our numerical examples indicate, the proposed methodology allows an efficient, principled and graphical exploration of different parameter combinations. The insights obtained from the model provide a “bird’s-eye view” perspective of different available options. After a set of options is selected, detailed and computational intensive simulations are used to choose the preferred one.



**Figure 7. Graphical views of optimization problem:** (a)  $c_c = L + 10C'$ , (b)  $c_c = 10L + 2C'$ .  $1/\mu = 1/\lambda = 8$  hours,  $S = 30$  users,  $1/\mu = 8$  hours,  $1/r = 120$  s,  $\tau = 0.01$  s,  $1/M = 1250$  Bytes,  $\Lambda = 900$ ,  $C_0 = 10$  Mbps. “Capacity” refers to  $C = C_0 + C'$ .

## 5. Discussion

The proposed methodology is a first step in the study of tradeoffs between statistical multiplexing and infrastructure costs in cloud systems. The analytical model introduced in this paper can be used to investigate the advantages and disadvantages of centralizing signaling servers. It allows for what-if analysis of different system parameters, and can be used to explore the state space in a principled way. In addition, it can also be used to assist practitioners in setting the state renewal timeout parameter ( $\tau$ ). Small values of  $\tau$  will yield frequent premature timeouts whereas larger values will delay the release of licenses of applications that unnecessarily remain active after a user leaves its desktop or after a crash.

We note that in order to obtain a tractable model, we made some simplifying assumptions some of which are discussed below.

**Channel model characteristics:** the channel model is one of the building blocks of our framework. In this paper, we consider an M/M/1 queue to model the channel characteristics. This model can be easily adjusted and adapted according to the needs, while still maintaining the general framework.

**Network protocol influence:** we do not model specifics of network protocols such as retransmissions or packet prioritization. Instead, we take a simplifying approach according to which a timeout occurs if a renewal packet experiences delay larger than the threshold  $\tau$ . The adjustment of the delay-related metrics according to different system characteristics is subject for future work.

**Available network infrastructure:** we assume an enterprise scenario in which publicly available cloud infrastructures cannot be used due to privacy and safety issues. Therefore, the infrastructure must be provisioned and planned by a single authority, which motivates the tradeoffs between infrastructure costs and multiplexing benefits discussed in this paper.

## 6. Conclusions

Cloud services are increasingly deployed across multiple and geographically distant sites creating a demand for a holistic performance evaluation before planning a converged computing-networking cloud service. We analyzed a case study inspired by a real-world oil and gas industry where a floating license service can be distributed among multiple data centers or centralized in a single pool. The centralized case is an example of network-computing cloud service in which the statistical multiplexing advantages of centralization can be overcome by the corresponding increase in communication infrastructure costs.

We derived an analytical model to evaluate tradeoffs in terms of application requirements, usage patterns and communication costs. The numerical results showed that the best solution depends on the relation of these several parameters. We believe that this model can serve as a guideline for capacity planning of computing and networks resources of floating licensing applications and can be a starting point for bridging the computing and networks aspects in an integrated manner.

## References

- [1] I. Foster, Y. Zhao, I. Raicu, and S. Lu, "Cloud computing and grid computing 360-degree compared," in *Grid Computing Environments Workshop, 2008. GCE'08.* Ieee, 2008, pp. 1–10.
- [2] P. Mell and T. Grance, "The nist definition of cloud computing," 2011.
- [3] S. Secci and S. Murugesan, "Cloud networks: Enhancing performance and resiliency," *Computer*, vol. 47, no. 10, pp. 82–85, Oct 2014.
- [4] Q. Duan, Y. Yan, and A. V. Vasilakos, "A survey on service-oriented network virtualization toward convergence of networking and cloud computing," *Network and Service Management, IEEE Transactions on*, vol. 9, no. 4, pp. 373–392, 2012.
- [5] M. Hamadani and W. Huffman, "Automatic software license manager," Apr. 21 1998, uS Patent 5,742,757.
- [6] V. B. Iversen, *Teletraffic Engineering Handbook.* ITU-D SG 2 and ITC, (New York NY), 2001, vol. 1.
- [7] P. Ji, Z. Ge, J. Kurose, and D. Towsley, "A comparison of hard-state and soft-state signaling protocols," in *Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications.* ACM, 2003, pp. 251–262.
- [8] L. Brandwacht, E. Meeuwissen, H. van den Berg, and M. Živkovic, "Models and guidelines for dimensioning private clouds," in *International Conference on Cloud Computing*, 2013, pp. 880–886.
- [9] H. Khazaei, J. Mistic, and V. B. Mistic, "Performance analysis of cloud computing centers using m/g/m/m+ r queuing systems," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 23, no. 5, pp. 936–943, 2012.
- [10] Q. Duan and Z. Zheng, "Holistic modeling and performance evaluation for converged network-cloud service provisioning," in *Advanced Information Networking and Applications (AINA), International Conference on.* IEEE, 2014, pp. 978–984.

- [11] L. Lewin-Eytan, J. Naor, R. Cohen, and D. Raz, “Near optimal placement of virtual network functions,” in *INFOCOM*, 2015.
- [12] L. Kleinrock, *Queuing Systems*. John Wiley & Sons Publishers, (New York NY), 1975, vol. 1.
- [13] M. Pióro and D. Medhi, *Routing, flow, and capacity design in communication and computer networks*. Elsevier, 2004.